D0 Computing Review

1) Networks
2) FNAL Farms
3) FNAL Analysis Facilities

Michael Diesburg D0/CD

Network Infrastructure

- Current network setup has several chokepoints
  - FCC <-> DAB (D0 Assembly Building)
    - Currently 3 pair of fibers
      - 1 for offline, 1 for online, 1 spare
    - Adding 6 more pair this year
    - Assume all can be driven at full Gb rate, will give us total of 7Gb/s bandwidth for offline.
    - Cost: ~$30K
  - DAB <-> Portakamps
    - Single Gb connection between each of the out buildings and switch in DAB.
    - Have fiber available to double this by adding Gb connections at switch in DAB
    - Cost: ~$8K

Network Infrastructure

- DAB Internal
  - Nodes in DAB are connected in groups of ~12-16 to hubs with 100Mb uplinks to 6509 switch.
  - Has already caused problems with network congestion when attempting to use nodes for analysis
  - Need to replace hubs with switches.
  - Should be done ASAP
  - Need to replace 21 hubs
  - Cost: ~$80K

Network Infrastructure

- Switch Upgrades: More connections needed
  - 6509 at DAB essentially full.
  - Will need to add connections for ~100 batch nodes on clued0 cluster plus additional Gb connections.
  - Cost 6509+blades: ~$100K
  - 6509 switch can handle max 384 100Mb connections
    - Will need at least 1 additional switch plus blades to fill out existing farm switch in FCC
  - Cost: ~$100K + $60K

Network Infrastructure

- Longer Term Plans:
  - Eventually will replace 1Gb backbone between DAB and FCC with 10Gb connections.
  - Has to drop significantly in cost before it is a reasonable thing to do (~2-3years?).
  - Will need to replace associated hardware on both ends to handle higher bandwidth.
  - Cost: ~$200K

Network Infrastructure

- Longer Term Plans:
  - Desktops currently have 100Mb connections
  - Expect bulk of traffic between DAB and FCC to be going to dedicated batch and server nodes.
  - However, demands of desktop could overwhelm 100Mb connections in a few years.
  - Running Gb to desktops would require major improvements to wiring and wholesale replacement of switch infrastructure.
  - Cost: ~$400K
  - Needs 10Gb backbone first
  - No decision to support copper Gb to desktop yet.

Network Infrastructure

- Offsite Connectivity
  - Effective use of resources at remote institutions for analysis and reprocessing is highly dependent on available bandwidth out of FNAL.
  - Current connections are OC3, to be upgraded to OC12 within the next year.
  - But connection is to ESNet. Most of our collaborators do not have direct connections to ESNet.
  - Better option is OC48 connection to STARLIGHT
  - Looking for source of dark fiber to use for connection
  - Assuming 1/3 bandwidth available to D0, would have ~6, 25, or 100MB/s with OC3,12,48 respectively.
  - We consider the STARLIGHT connection crucial to effectively using remote resources.

FNAL Farms

- Current configuration:
  - 122 dual processor nodes (500,750,1000MHz)
  - 8-processor O2000 w/1TB buffer for I/O
    - Separate dedicated Gb connections to Enstore and worker
  - .186THz equivalent compute power
  - Capable of reconstructing 12Hz with current code
  - Efficiency in 70-80% range.
  - Adding 128 2GHz nodes soon
    - Will bring capacity to .565THz

FNAL Farms

- Most cost effective plan for IIb farms would be to buy full installation in FY05 at beginning of run.
  - Would require > $1.0M in FY05
  - Would leave us with 2 year gap with no farm upgrades.
- Will need to phase in purchase over FY03, 04, and 05 with bulk of purchase as close to run as possible.
- Base assumptions for estimating farm size:
  - 50 Hz average DAQ rate
  - 50 sec processing time on 500MHz equivalent cpu
  - 70% farm production efficiency
  - 60% additional cpu needed for re-processing, other tasks.
  - $2500 cost for dual cpu node with 1GB memory and disk
  - Spending profile of 20, 30, 50% in FY03,04,05 respectively
  - 3, 4, and 6GHz cpus available in FY03,04,05 respectively
  - $25K I/O capacity cost per 100 worker nodes

FNAL Farms

- Note on timing assumptions:
  - There are large uncertainties in processing time
  - Preliminary test show ~2X increase to next production version
  - Changes in trigger mix can significantly effect average complexity of events to be processed.
  - Accelerator configuration has major impact on processing times due to difference in number of interactions/crossing
    - e.g. estimated 32 vs 80secs/event for 132 and 396ns crossings respectively

Average Rate:	50
Farm Efficiency:	70%
Misc. Processing:	10%
Reprocessing:	50%
Cost/node:	2,500
I/O Cost/100 nodes	25,000

CPU	SpecI2000
3GHz	960
4GHz	1280
6GHz	1920
10GHz	3200
15GHz	4800

FY05 Target Spending Fraction		20%		30%		50%		Total	
Execution Time	THz CPUs at Beginning of Run	FY03, 3GHz Nodes		FY04, 4GHz Nodes		FY05, 6GHz Nodes		Target	
		No. Nodes	Cost	No. Nodes	Cost	No. Nodes	Cost	No. Nodes	Cost
50	2.9	80	225,000	120	350,000	200	575,000	400	1,150,000
32	1.8	51	152,500	77	217,500	128	370,000	256	740,000
80	4.6	128	370,000	192	530,000	321	902,500	641	1,802,500
120	6.9	192	530,000	289	797,500	482	1,330,000	963	2,657,500

50sec:	Nominal timing for planning purposes
32sec:	Best case scenario with 132ns crossing (unlikely)
80sec:	Best case scenario with 396ns crossing
120sec:	Possible case with 396ns crossing

FNAL Analysis Facilities

- Are implementing 3-tiered approach to analysis facilities based on size of data sets:
  - Large central analysis - ~10TB data sets
  - Small batch analysis - ~1TB data sets
  - Desktop analysis - ~100GB data sets
- Currently in place or in construction:
  - CLueD0 - desktop analysis cluster
  - CLuB - ClueD0 backend
  - D0mino and CAB - SGI 02000 + central analysis backend

FNAL Analysis Facilities

- D0mino
  - 176 300MHz processors, 88GB memory, 35TB disk, 8 Gb interfaces.
  - Strengths:
    - Large interactive load, typically ~150 users, load average ~80.
    - Large storage capacity and I/O rate, typically ~5TB/day at present
    - Large network I/O capacity.
    - Stable OS, requires relatively little management attention.
    - Good maintenance available.

FNAL Analysis Facilities

- D0mino
  - Weaknesses:
    - Cost
    - Relatively slow processors due to age. Too expensive to upgrade.
    - Maintenance is expensive and will likely increase as system ages further.
  - Maintenance costs will dictate that we decommision system in ~2years.
  - Have not resolved issue of how to replace functionality.
  - Linux based SMP systems not ready for prime time yet.

FNAL Analysis Facilities

- CAB - Central Analysis Backend
  - Attempt to migrate compute intensive processes from D0mino to faster, cheaper commodity processors.
  - 16 Linux dual processor prototype systems
  - Uses PBS for batch submission from D0mino
  - SAM will be used for data access both from tape and from D0mino cache.
  - Dedicated Gb data path from D0mino
  - Will add 128 2-GHz processors this summer (same bid/order as farm upgrade).

FNAL Analysis Facilities

- CLueD0
  - Cluster Linux Environment at D0.
  - Foremost a desktop cluster, but also provides code development platform and processing farm
  - An institute based cluster, i.e. institutes provide both hardware and management manpower.
  - Management shared between ~20 part time admins.
  - Resources partially allocated on basis of contributions.
  - Has ~200 batch queue slots for processing small datasets (~100GB) with limited network I/O requirements

FNAL Analysis Facilities

- CluB - ClueD0 Backend
  - Similar to CAB but geared toward smaller (~1TB datasets).
  - Will have commodity disk server.
  - Located in DAB for direct connection to network switch.
  - Will be managed and built by institutes the same way CLueD0 is.
  - SAM will be used for access to tape resident data
  - Data delivery from FCC will be over dedicated data network link

FNAL Analysis Facilities

- CLueD0-CLuB
  - Administration is both a strength and weakness of these systems.
  - Distributing effort allows us to operate a cluster we would not otherwise have manpower to support.
  - But continuity of management is a problem as grad students and post-docs move on to other things.
  - 24x7 support isn't practical
  - Would benefit greatly by having a full-time professional administrator to oversee volunteer sysadmins.

Conclusions

- Need to augment current chokepoints in network at D0 and between D0 and FCC
- Improvement on FNAL off-site connectivity is crucial to effectively using remote resources
- Farm expansion to meet needs is straight forward, but large uncertainties exist in amount of compute power needed
- Exploring ways to augment compute power of central systems with commodity processors
- Institute contributions to CLueD0-CLub will be significant part of computing resources.